Integration of multi-sensorial stimuli and multi-modal interaction in a hybrid 3DTV system

FRANCISCO PEDRO LUQUE, IRIS GALLOSO AND CLAUDIO FEIJOO, Center for Smart Environments and Energy Efficiency (CEDINT), Universidad Politécnica de Madrid
CARLOS ALBERTO MARTÍN AND GUILLERMO CISNEROS, Grupo de Aplicación de Telecomunicaciones Visuales (G@TV), Universidad Politécnica de Madrid

This paper proposes the integration of multi-sensorial stimuli and multi-modal interaction components into a sports multimedia asset under two dimensions: *immersion* and *interaction*. The first dimension comprises a binaural audio system and a set of sensory effects synchronized with the audiovisual content, whereas the second explores interaction through the insertion of interactive 3D objects into the main screen and on-demand presentation of additional information in a second touchscreen. We present an end-to-end solution integrating these components into a hybrid (internet-broadcast) television system using current 3DTV standards. Results from an experimental study analyzing the perceived quality of these stimuli and their influence in the Quality of Experience are presented.

Categories and Subject Descriptors: **H.5.1 [Information Interfaces and Presentation]**: Multimedia Information Systems—*Video*; **H.5.2 [Information Interfaces and Presentation]**: User Interfaces—*Interaction styles;* **I.4.9 [Image Processing and Computer Vision]**: Applications; **J.7 [Computers in Other Systems]**: Consumer products; **H.1.2 [Models and Principles]**: User/Machine Systems—*Human Factors***.**

General Terms: Multimedia Information Systems

Additional Key Words and Phrases: multi-sensorial multi-modal media, immersive media, interactive media, hybrid-based 3DTV, binaural audio, sensory effects, interactive 3D objects integrated into the video scene, second screen, quality of experience.

ACM Reference Format:

Francisco Pedro Luque, Iris Galloso, Carlos Alberto Martín, Claudio Feijoo and Guillermo Cisneros. 2014. Integration of multi-sensorial stimuli and multi-modal interaction in a hybrid 3DTV system. ACM Trans. Multimedia Comp. Commun. Appl. N, N Article (Month YYYY), 20 pages.

DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Quality of Experience (QoE) studies in multimedia applications encourage the integration of multi-sensorial stimuli as a mean to enhance the user experience by stimulating senses other than sight and hearing [Le Callet et al. 2012]. This line of research is supported by several previous studies revealing the positive influence of spatial audio, ambient lighting and olfactory stimuli among other effects, in the perceived quality of a multimedia asset [de Ruyter and Aarts 2004; Ghinea and Ademoye 2012], particularly for specific genres such as documentaries, sports, and action movies [Waltl et al. 2010]. The recently launched ISO/IEC 23005 Information technology – Media context and control standard (MPEG-V) represents an important step in the consolidation of this new approach to multimedia

The work described in this paper has been developed partially within the framework of the *ImmersiveTV* project funded by the Spanish Ministry of Industry Tourism and Commerce Grant #TSI-020302-2010-61.
Authors' addresses: F. P. Luque, I.Galloso, and C.Feijoo, Edificio CeDInt-UPM, Campus de Montegancedo, 28223, Pozuelo de Alarcón, Madrid, Spain; email: [franluque,iris,cfeijoo]@cedint.upm.es; C. A. Martín and G. Cisneros, GATV ETSI de Telecomunicación, Ciudad Universitaria s/n, 28040, Madrid, Spain; email: [cam,gcp]@gatv.ssr.upm.es.

DOI:http://dx.doi.org/10.1145/0000000.0000000

consumption. However, many questions remain concerning not only the specific effects that can better complement a given audiovisual content, but also the best way in which to integrate and combine them to enhance the QoE.

The history of television broadcasting is full of attempts to integrate interaction in order to enrich the multimedia experience with new features, functionalities, and/or information. However, even when surveys have found that users consider interactive television services innovative and desirable [INSPIRA 2006], very few applications have shown clear added value for them in practice. The interactive TV concept is nowadays integrated in the new Connected TV or Hybrid TV paradigm. Despite the fact that Hybrid Broadcast Broadband TV (HbbTV) is not completely deployed, services available in European countries show the importance of video on demand applications, taking advantage of the broadband network to deliver personalized audiovisual content, as catch-up applications. Although these initial implementations demonstrate the practical feasibility of the approach, further efforts are required to overcome the technical challenges associated with the wide deployment and adoption of HbbTV. Furthermore, the combination of sensory effects with interaction components within a multimedia experience remains unexplored practically.

The work presented in this paper aims to complement the main audiovisual content of a sports 3DTV multimedia asset with a set of multi-sensorial stimuli and multi-modal interaction components aimed to enhance the QoE. According to the nature of their contribution to the user experience, the elements introduced are classified under one of two dimensions: immersion or interaction. Under the immersion dimension, we group those additional elements identified in previous works as contributing to the subjective feeling of being part of the displayed environment. In our approach, these components are those that stimulate the aural, olfactory, and visual (in particular the peripheral vision) senses synchronously with the main audiovisual content. The interaction dimension comprises those components that promote a proactive, action-driven involvement of the user in the experience. Therefore, it is aimed at enhancing user control over the experience through interaction.

We have developed an end-to-end solution enabling the seamless integration of the aforementioned stimuli and interaction components into a hybrid (internet-broadcast)-based television system. Our aim is to evaluate, from a practical perspective, the technical feasibility of this concept with regard to the 3DTV standards currently available.

The remainder of this paper is organized as follows. Section 2 presents an overview of relevant work with regard to our approach and objectives. Section 3 introduces the conceptual framework of our work. Section 4 presents the architecture of our technical solution as well as the mechanisms and strategies adopted across the entire workflow. Section 5 discusses the preliminary results from an experimental study analyzing the influence of the introduced multi-sensorial stimuli and interaction components in the QoE. In Section 6, we discuss the results of our work and highlight open issues and research challenges remaining to be tackled in the future.

2. RELATED WORK

2.1 Immersive Media

Immersion has been studied widely in the context of virtual environments and multimedia applications (a comprehensive overview with a focus on virtual environments can be found in [Schuemie et al. 2001]). In our work, we adopt the definition provided in Slater and Wilbur [1997], where immersion is defined as the perception mechanism dealing with the user's mental state of feeling the relationship between self-awareness and the surrounding environment. In this sense, we can say that the state of full immersion is achieved when all the senses of the human body are engaged in the multimedia experience.

Considering the previous definition, immersive media deals with the creation of multimedia content produced to influence human senses in a particular way. In the context of television, vision

and hearing have been addressed more commonly over other senses, as they are considered as having the greater influence on the feeling of presence. Regarding the sense of sight, the best way to create immersion has been to provide users with the capability of watching 3D content. This has been exploited during the last decade thanks to the recent development of 3D technologies and research in the capture, production, and delivery of stereoscopic content. Some works have proposed the use of 3D environments to offer immersion in tele-presence applications, such as multi-party communications [Yang et al. 2010]. Traditionally, hearing has been enhanced by means of multi-channel surround audio monitors located around the listener.

Investigations into different implementations of immersive media to improve the user sensorial experience can be found in the literature. In the case of vision, the works of Jones et al. [2013] and Mills et al. [2011] propose a proof-concept system to augment the area around the television screen with the projection of light visualizations in the periphery that complement the content being displayed on screen. The main goal of this system is to increase the field of view and to create an atmosphere that envelops the user. A similar approach is followed by the *Philips' Ambilight* TV screen [Weffers-Albu et al. 2011] in which the background light of the television display is adjusted in real time based upon the picture content to create different effects.

Conversely, the senses of smell and touch have been considered rarely, mainly because they are substantially more challenging to implement in a realistic and effective way. Within the few examples available, the works by Ghinea and Ademoye [2012] and [Murray et al. 2013] study the impact of olfactory effects in the QoE and explore users' tolerance to synchronization errors between olfactory data and the audiovisual content. Likewise, Cha et al. [2009] propose a touchable 3D video system in which users are able to feel a video scene through a force feedback device.

Aware of the increasing influence of multi-modal and multi-sensorial media in the immersive television experience, the Moving Picture Expert Group (MPEG) has been working since 2009 on the definition of the MPEG-V standard. This standard provides an architecture and associated information representations (metadata) for the interaction and interoperability between virtual worlds (i.e., digital content) and real worlds through various sensors and actuators. Part 3: "Sensory Information" of the standard [ISO/IEC 2013] defines a set of sensory effects (e.g., light, temperature, wind, vibration, touch) as well as semantics that the content creator may use to deliver multi-sensorial content in association with audiovisual data. Waltl et al. [2013] and Yoon [2013] demonstrated an effective end-to-end framework implementation for the creation and delivery of multi-sensorial data synchronized with audiovisual content using the MPEG-V standard.

2.2 Interactive Media

The history of TV broadcasting is characterized by multiple attempts to achieve a more interactive viewing experience. These attempts have tried to provide a more active role for the viewer, prompting them to be no longer a passive consumer of information. The first broadly deployed interactive system was Teletext, specified in Europe in the 1970s [Heightman 1975]. Teletext was an intrinsically digital system that took advantage of the analog signal characteristics to supply additional information for the viewers in the form of an electronic newspaper. However, Teletext suffered two severe limitations: the reduced bitrate due to its insertion in the vertical blanking interval of the analog signal, and the lack of a return channel. This kind of interactivity is called local because the user can only choose among the pieces of content that are available locally, which results in a very limited system.

The advent of digital TV, first using satellites as the means of transmission, followed by cable networks and finally, terrestrial and digital subscriber line networks, brought new opportunities for interactive TV due to the flexibility of digital information to integrate new services and content.

Proprietary specifications were developed (such as OpenTV and MediaHighway) to provide interactive services in vertical markets (i.e., pay TV platforms). Moreover, the Digital Video Broadcasting project (DVB) created a specification of standard interactivity middleware for horizontal

markets (e.g., free-to-air digital terrestrial TV): MHP – Multimedia Home Platform [ETSI 2006]. For a variety of reasons, MHP did not actually deploy (except in a few countries such as Italy) and it is now considered a failed standard. More success has achieved MHEG-5 [ISO/IEC 1998], an ISO standard adopted in the United Kingdom to replace the Teletext service.

Although surveys with real users reveal that interactivity is appreciated by the viewers [INSPIRA 2006], very few applications have shown clear added value for them. This fact explains the failure of deployment of interactive TV systems in the past, despite the expectations of TV broadcasters, content providers, and engineers. As shown in the report of the INSPIRA project [2006], even though users may consider that the availability of interactive services on TV is innovative and desirable, there is no actual demand for this kind of content.

The interactive TV concept is nowadays integrated in the new Connected TV or Hybrid TV paradigm, characterized by TV sets able to receive and to play content coming from both broadcasting networks (like digital terrestrial television) and the Internet (i.e., the broadband network). In this way, Connected TV avoids the weak point that has plagued other interactive TV systems; the lack of a return channel. Moreover, the HbbTV initiative has created a standard specification that allows both manufacturers and TV broadcasters to exploit their content portals [ETSI 2012b]. This standard references a previous DVB norm for the signalling and carriage of the interactive applications [ETSI 2010b]. Some authors have proposed to use Internet TV to provide personalized programmes such as news [Olsen et al. 2012].

Our development efforts concerning interaction focus on overcoming the technical challenges associated with the implementation of the multi-modal interaction components considered and their integration in the 3DTV transmission chain. As described further in Section 4, we rely on a signalling mechanism that uses the *application information table*, defined by DVB [ETSI 2010b] for the delivery of additional content, in order to observe the backward compatibility restriction. This mechanism is used currently by TV broadcasters in Connected TV deployments to signal their HbbTV portals in the broadcast stream. However, the additional content delivered by most of these deployments is limited to multimedia and HTML (or CE-HTML) pages.

2.3 Enhancing user experience in multimedia applications

User experience in multimedia applications is a live and fertile research field aimed at gaining insight into the factors and mechanisms that influence the subjective quality assessment of a multimedia asset (i.e., the content quality as perceived by an individual). The study of these phenomena has been encompassed into the concept of 'Quality of Experience', which is defined as *"the degree of delight or annoyance of the user of an application or service"* [Le Callet et al. 2012].

The QoE has been found influenced by a combination of interrelated factors of contextual, technical, and human nature. Contextual factors have been defined by Jumisko-Pyykkö [2011] as those *"that embrace any situational property to describe the user's environment"*. These not only concern the physical context, but also other dynamic or static features of economic, social or technical nature, including other concurrent activities in which the user can be involved [Le Callet et al. 2012]. Technical factors (also known as system factors) refer to those conditioning the resulting technical quality of an application or service [Jumisko-Pyykkö 2011]. Different categories of technical factors have been proposed in the literature, both from a technical perspective, in which they are divided according to the related components of the service architecture/chain (e.g., in Le Callet et al. [2012]), and from a user perspective, considering their final influence/manifestation during the multimedia experience [Bracken et al. 2011]. Finally, human factors comprise those features that characterize the user and have an influence on his/her perception of quality. In the context of media, the quality perception mechanism is stimulated on two principal levels: the early sensory processing level, and the high-level cognitive processing enabling conscious interpretation and judgement [Goldstein 2010; Jumisko-Pyykkö 2011]. Quality perception mechanisms, as well as those enabling media enjoyment,

have been analysed across a great variety of genres as a dependent variable of personality traits, individual differences, mood, content characteristics, social context, or as a combination of these [Raney and Bryant 2002; Slater 2003; Zillmann 2003; Nabi and Krcmar 2004; Wechsung et al. 2011].

In our work, we are particularly interested in analysing how the quality perception mechanisms can be influenced positively by four specific types of low-level and high-level stimuli (and by a combination of them). This line of research is supported by several previous QoE studies proposing the integration of sensory stimuli (beyond the conventional audiovisual content) into a multimedia experience in order to strenghen immersion. For instance, de Ruyter and Aarts [2004] implemented and tested Ambient Intelligence scenarios aimed at complementing audiovisual content with meaningful lighting effects. In a broader approach, Timmerer et al. [2012] introduced sensory effects that are rendered on complementary devices in synchronization with the actual multimedia asset. In one of their experiments, the authors analyzed the influence of wind, vibration, and light effects on the user experience across different genres, including action movies, sports, news, documentary, and commercials [Waltl et al. 2010]. They found that the user experience with documentary, sports, and action genres was influenced positively by the introduction of sensory effects.

3. CONCEPTUAL APPROACH OF IMMERSIVE TV

In our work, we aim to complement the main audiovisual content of a sports 3DTV multimedia asset (in our case study, a football match) with a set of multi-sensory stimuli and multi-modal components. Fig. 1 illustrates the concept, general architecture, and principal elements of our approach. Considering their nature, these are classified under one of two dimensions: *immersion* or *interaction*.

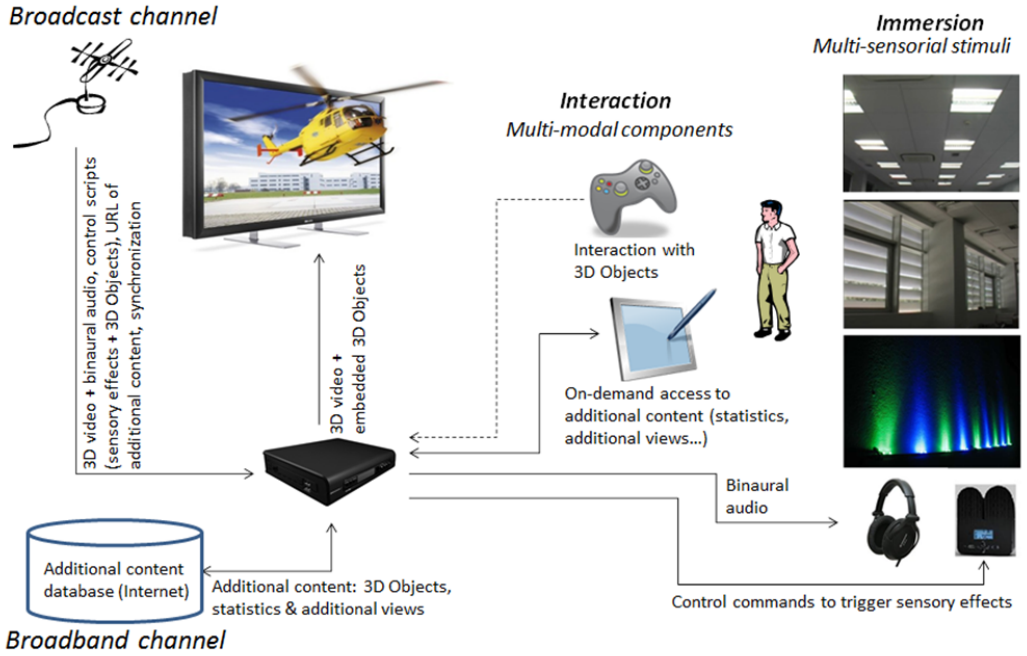

Fig. 1 Conceptual framework, general architecture, and main components of our approach

Under the *immersion* dimension, we group those additional elements identified in previous works as contributing to the subjective feeling of being part of the displayed environment. In our approach, these components are:

(i) a binaural audio system simulating the spatial distribution of the audio sources in the displayed environment (auditory stimuli),
(ii) a set of sensory effects rendered –in synchronization with the main audiovisual content– on specific devices, comprising an electronic scent vaporizer (olfactory stimuli), lighting, and shutter controllers (peripheral vision stimuli).

The human auditory system is capable of perceiving the surrounding environment even more acutely than the vision system, because it is not limited by the direction from which it receives the information (sounds can be perceived from behind the listener and at different heights). The 3D binaural audio, introduced as auditory stimuli in our case study, simulates the acoustic environment of the football stadium using virtual audio sources. The resulting audio signals are reproduced using specialized headsets to evoke the feeling of being physically surrounded by the audio sources.

In response to specific events registered during the football match, different actuator devices are triggered synchronously to deliver a combination of sensorial stimuli designed to extend the perception of the displayed scenes beyond the main audiovisual content. Table 1 summarizes the actuator devices used to deliver the sensory effects and the associated triggering events.

Table 1. Actuator devices, related sensory effects, and corresponding content events

Actuator device	**Sensory effect**	**Content event**
Shutter control	The room blinds open or close	Start/ending of the match
Fluorescent lighting control ballast	The lights are turned on or off	Start/ending of the match
Led-based ambient lighting system	Ambient light color and intensity is controlled to create different illumination atmospheres	Start/ending of the match Strategic moments of the game
Scent vaporizer	Cut grass scent is vaporized	Strategic moments of the game

The *interaction* dimension comprises those components promoting a proactive, action-driven involvement of the user in the experience, i.e., it is aimed at enhancing user control over the experience by means of interaction. The components under the *interaction* dimension are:

(i) interactive 3D objects related to the main audiovisual content, which are rendered in synchronization with and integrated into the 3D scene displayed through the main screen,
(ii) additional information such as statistics and additional views presented on-demand using a portable tactile device (i.e., secondary screen).

The idea of enriching the three-dimensional video content through the integration of interactive computer-generated objects into the video scene has been inspired by Augmented Reality applications. The aim is to add meaningful information (e.g., in the form of 3D objects, visual effects, and text annotations) capable of enhancing the user experience and of providing added value to the content provider at the same time. At specific (synchronized) moments of the match, especially those in which the gameplay is stopped (e.g., breaks and replays), the user is given the opportunity to interact with superimposed computer generated objects using a joystick device.

Finally, the user is given the opportunity to access additional content through a second screen (we have used a tablet-pc, but any device capable of showing web content would be suitable). In our case study, the additional content consists of real-time data, statistics, and multimedia content (additional views) of the football match event.

Considering all these elements, the experience from the user perspective flows as follows. As soon as the football match starts, the room is automatically configured to enhance the viewing conditions. The fluorescent lights are turned off and the shutter blinds are closed. The ambient lighting is turned on and its color is regulated to create a calm and relaxing atmosphere. If binaural audio is activated, the user can perceive the spatial distribution of audio sources as if he/she were just in the field. Cut

grass scent is vaporized at strategic moments of the game to reinforce the feeling of immersion. During the match, the user can access replays and updated statistics regarding the game and the players at any time using a tablet device. He/she can also interact, using a standard joystick, with meaningful 3D objects that are integrated into the 3D video scene during the game and advert breaks.

The table in Appendix A.1 summarizes the type of content, synchronization requirements, and rendering/visualization terminal corresponding to each content category presented in our testbed. These guidelines have been fed as inputs to the design and implementation phases to guarantee a coherent presentation of the immersion and interaction components during the experience.

4. IMPLEMENTATION

To guarantee the compatibility of our solution with current broadcast frameworks, our design focuses on the adaptation of the conventional transmission chain to support the additional interactive services and sensory effects of the case study. In particular, our implementation fulfills the DVB standards for television broadcasting: the 3D video satisfies the DVB-3DTV specification and a professional DVB-T modulator has been used to emulate a real DTV transmission. The coding and multiplexing of additional data are also based on standard technology: the DVB norm for hybrid broadcasting (the same norm specified by HbbTV, the Connected TV standard) and MPEG-2 private data packets.

In the related work section, the release of MPEG-V and the possibility of using this standard to build an end-to-end broadcasting transmission system with sensory information was noted [Yoon 2013]. Even though the specification phase of our project was completed before the official release of MPEG-V part 3, our solution shares several key aspects with the standard. The most important of these are the inclusion of sensory effect metadata in the media stream and the use of a Media Processing Engine (called in the project "Receiver Gateway") to manage and deliver the different kinds of data. For this reason, our proposal can be easily migrated to fulfill the MPEG-V standard. This migration would require to create the *Sensory Effect Metadata* according to the *Sensory Effect Description Language* (SEDL) and the *Sensory Effect Vocabulary* (SEV) specified by the standard, considering the specific devices used in our test bed (KNX sensorial devices). Likewise, the receiver gateway would require some minor changes to generate the suitable commands for the actuators.

Fig. 2 represents the high level architecture of our solution emphasizing the design options adopted at each stage of the transmission chain. Throughout the architecture, and depending on the type of multimedia content, the data are transmitted using either the traditional unidirectional broadband framework or the bidirectional broadcast network. For each case, the flow of transmission is depicted using blue and red arrows, respectively. Our architecture solution exploits the broadcast framework to deliver the main audiovisual content as well as the additional multi-sensorial stimuli and multi-modal interaction that have to be presented in synchronization with the main signal. On the other hand, the broadband network is used to transmit the on-demand content intended for the *second screen* and the 3D objects geometry that can be demanding considering their size. In the final stage, all the content is delivered using the viewer's private IP network that connects the visualization terminals and sensorial devices. The depicted transmission flow is used as a reference to describe the details of our implementation throughout the present section.

4.1 Content creation and adaptation

In our test bed, the principal audiovisual content consists of a digital 720p stereoscopic video sequence of a premium football match between the Spanish teams F.C. Barcelona and Real Madrid C.F. In order to provide a high quality video source, the sporting event was recorded in *Digital Nonlinear Extensible High Definition* format at 100 Mbps. At the recording stage, a specific setup of Panasonic LDK 800 cameras with Canon optics was deployed covering the most important areas of the game field, as shown in Fig. 3. The cameras were mounted on stereoscopic rigs to produce a full

resolution signal for each eye in a side-by-side format. Therefore, the effective captured resolution is 2560 × 720 pixels, where half of the width is reserved for each separate view.

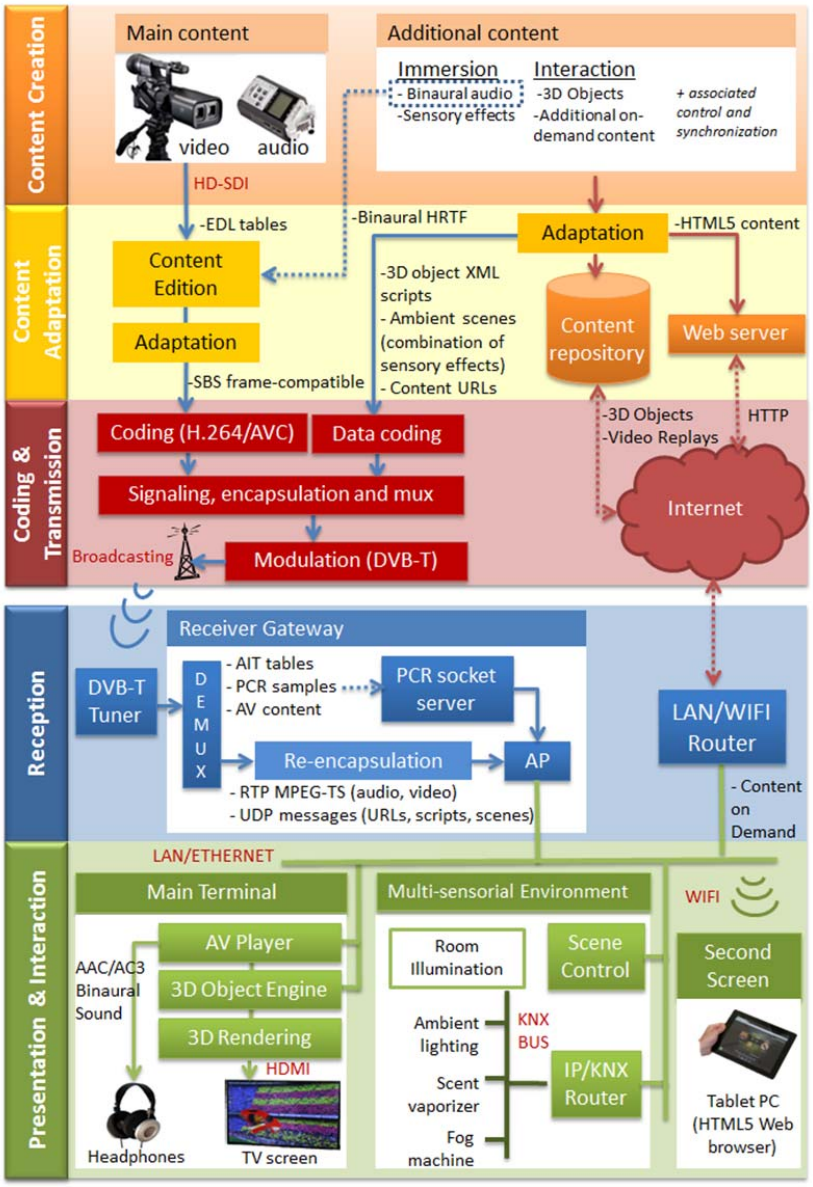

Fig. 2. Architecture of our end-to-end multi-sensorial and multi-modal solution

On the user's main visualization terminal, 3D objects are enabled to be displayed at specific moments of the match (Fig. 4). In addition to the object geometry, a set of behaviors and animations programmed to be played through the user's interaction are also defined at this stage. All these data are packed in one propietary file by the content provider to be transmitted via internet.

The second category of additional content aims to provide real-time information related to the main content over the *second screen* viewing terminal. This information comprises an event log with live commentaries, statistics about the football match and teams, and gameplay replays of specific video

shots from cameras different from those providing the main video signal. This kind of content has to be displayed in synchronization with the playing time of the match, such that it is not presented before the corresponding game event has taken place on the main screen. For this reason, the data are stored with a timestamp reference that is used to filter the content in the following stages.

Stereoscopic cameras
1- Master
2- Medium-shot (left)
3- Medium-shot (right)
4- Left goal
5- Right goal
6- Sideline Over/Under
7- Beauty shot
8- Steady pool
9- Offside left
10- Offside Right

Fig. 3 Spatial distribution of the camera rigs used to record the audiovisual content

The final type of additional content comprises the sensory effects script commands. In our implementation, these have been coded according to the format defined by the KNX protocol standard [ISO/IEC 2007] of the rendering devices being used. In this sense, each of the actions described in Table 1 are converted into an appropiate KNX frame that can be interpreted and excuted by the corresponding device at the reception stage.

After its creation, the raw content needs to be adapted and prepared for transmission through the corresponding delivery channel. With the final video production, the list of temporal references to synchronize the additional content can be specified and encapsulated together in the broadcast signal. In our test bed, these references are defined for the interactive 3D objects to be displayed on the main visualization screen and for the command scripts that trigger the rendering of sensory effects in the actuator devices integrated in the visualization room.

Fig. 4. Examples of low polygon 3D objects modeled for the football match scenario

Concerning the video content, an editing decision list is defined to select the most suitable sequences of the final cut (see table in Appendix A.2). In this process, the content length has been reduced to half an hour. After the edition, the audiovisual content is converted to a frame compatible side-by-side format. Among other possibilities, such as service-compatible and hybrid that were also considered in the project, the frame compatible format is found the most suitable with current industrial standards and particularly with the DVB-3DTV specification [ETSI 2012a]. Taking into consideration the original video resolution obtained in the previous stage, it is necessary to perform a subsampling step before encoding the video stream, having as a result a full-resolution interlaced signal (1920 × 1080 pixels) at 25 fps formed by two half-resolution views.

Together with the video, the recorded audio is also processed to produce a 3D binaural sound signal that can be used for transmission in a broadcast scenario [Cobos et al. 2013]. In particular, the algorithm used to adjust the audio frequencies is based in a parametric definition of the *head related transfer function* described in the work by Ramos and Cobos [2013]. In the presentation stage, the resulting binaural sound is delivered by means of isolated headphones.

The additional content also has to be adapted in order to prepare the data for its transmission and interpretation by the visualization terminals. The table in Appendix A.3 summarizes the different actions performed to convert the content into the desired formats, as well as the delivery channel used in each case. Once the contents have been adapted, the data that require higher bandwidth are stored in a web server repository (e.g., 3D objects geometry and video replays), whereas the other additional data are multiplexed in the broadcast stream signal. These may also include those events that need to be interpreted simultaneously in an actual broadcast scenario (e.g., a command to enable or disable a certain 3D object that has been downloaded previously via broadband).

In addition to representing the 3D elements on the screen, the multi-modal interaction defined in the scenario implies programming a group of actions or animations that can be played on these objects during the visualization. On the receiver side, these actions can be triggered either by the direct local interaction of the user (using a gamepad device) or remotely through script messages encapsulated in the broadcast stream. For the latter case, a mechanism of communication based on the XML remote procedure call specification (XML-RPC) [Winer 1999] has been defined (see Fig. 5). The format has been declared in a "document type definition" file so that messages can be parsed and interpreted correctly. Each sent message represents an *event* that has to identify three elements: a specific *object* to interact with, an *action* to be performed with corresponding attached *params*, and the *time* (in number of frames) that the receiver should wait to accomplish the process.

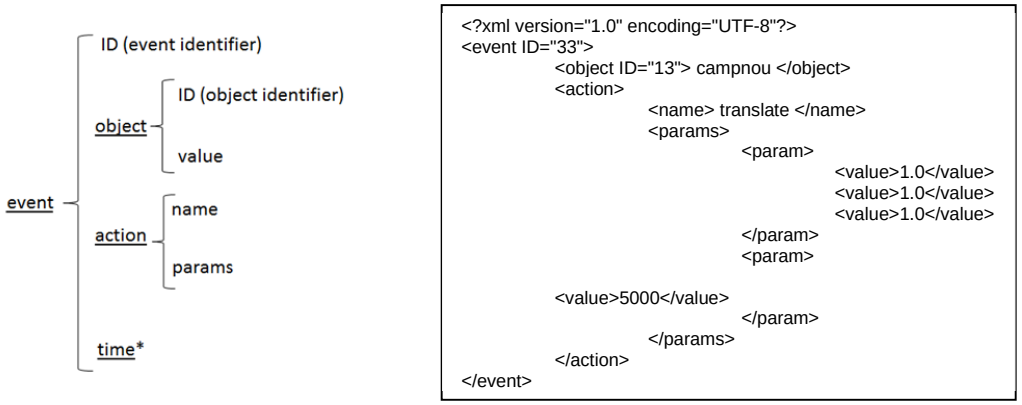

Fig. 5. (Left) Hierarchical representation of a valid constructed XML message; (Right) example of a command script to translate a 3D object

Finally, the command messages triggering the rendering of sensory effects are translated into a list of sensorial scenes to facilitate their transmission via broadcast. These scenes encode a set of KNX frame commands to be executed at the receiver side to complement meaningful events of the game with a specific combination of multi-sensorial stimuli (see Table in Appendix A.4). This approach has the advantage of coding all the command frames into only one text message, which reduces the amount of information that needs to be sent. This is of great importance in our architecture considering the limited bandwidth of any broadcast signal.

4.2 Coding and transmission

According to the DVB-3DTV specification, and to ensure compatibility with conventional HD receivers, the produced video was encoded in the AVC format at 11.3 Mbps. The following DVB-T parameters were selected for the transmission: bandwidth = 8 MHz, convolutional code rate = 1/2, modulation = 16 QAM and guard interval = 1/32. These parameters allow a total useful bitrate of 12.06 Mbps, which is sufficient to include the video, audio, and additional data. In our test bed, a DVB-T modulator configured with this setup was used showing that the selected bitrate and

parameters are appropriate for an actual broadcast scenario. In fact, an adequate selection of DVB-T parameters (most efficient modulation for the *Coded Orthogonal Frequency Division Multiplexing* carriers, least redundancy code rate, and shortest guard interval) would allow a transmission bitrate of 31.6 Mbps, considerably higher than the result value obtained for the previous setup.

A signaling table is included in the transmitted mux to inform the visualization terminals on the receiver side about the URL from which to download additional content. This is also the mechanism used in the Connected TV based on the HbbTV standard [ETSI 2012b] to provide content via the broadband network. The signaling table has been designed using the *application information table* specified by DVB for hybrid environments [ETSI 2010b] as a model.

On the other hand, command scripts addressing 3D objects and sensory devices are less demanding in terms of bitrate and therefore can be encapsulated into the broadcast transport stream. For this purpose, we use packetized elementary stream packets containing private data, according to the MPEG-2 specification – system layer [ISO/IEC 2000]. The overhead caused by these messages is not significant: the complete MPEG-2 transport stream comprises around 17 million packets, of which 71 TS packets are used for the XML command scripts and 12 TS packets are dedicated to encode the multi-sensorial scene messages.

As some 3D object animations and all sensory effects must be activated in synchronization with the main audiovisual content, timing information is also needed. This requirement has been fulfilled by using presentation time stamps referred to the Program Clock Reference (PCR), such that an accurate synchronization is guaranteed between the audiovisual content and the additional effects. During the implementation tests, a maximum delay of around one second was measured between a specified timestamp and the triggering of the corresponding sensory effect in the rendering device.

4.3 Reception

The main goal of the reception stage is to process the multiplexed signal and to deliver the different content in the correct format to the corresponding user terminals involved in the presentation. In our architecture, these terminals are connected together through a local area network (LAN) to the receiver to support a direct communication between them and to guarantee the scalability of the solution. Depending on the type of content, the manner in which information is delivered to the IP network varies.

With regard to the audiovisual content, the process of re-encapsulating MPEG-TS frames to be retransmitted through the local IP network consists of extracting the different DVB services from the mux signal and creating a new MPEG-TS frame that contains only the extracted service. In a local IP network, as we have configured in our architecture, the mechanism used to transmit DVB services is defined by the DVB-HN standard [ETSI 2010a], based at the same time in the real time protocol (RTP) for the transmission [Schulzrinne et al. 2003]. This protocol includes timestamps related to the PCR in the MPEG-TS frame that are used to estimate and fix the *jitter* introduced by the network and to synchronize the *time drift* between the transmitter and the receiver. The receiver sends the RTP packets using a multicast IP address, such that the audiovisual content can be played simultaneously on the different terminals connected to the network.

In the case of additional content, we have followed the approach specified by HbbTV: an application information-signaling table is included in the broadcast stream to specify the URL at which content can be found via Internet. At processing time, the receiver has to extract this table from the stream, parse its information, and make the parameters available to the viewing terminals through the LAN network. To notify the availability of the content, a *user datagram protocol* message is sent using a multicast IP address; thus, depending of the type of content, each terminal can decide whether to process or not the data.

However, there is a different approach to accessing the content to be shown on the *second screen.* In this case, a socket service has been included in the receiver gateway such that the terminals can

establish a connection to synchronize with the PCR samples that have been extracted during the demultiplexing process of the audiovisual content. These samples are used by the second screen to access the remote web server through an HTML browser such that the information provided by the server is synchronized with the content displayed on the main screen whenever the setup is launched.

4.4 Presentation & Interaction

The final module in the architecture includes the devices for the end user visualization of the multimedia content (TV signal and additional content), as well as the set of actuator devices rendering the sensory effects and the accessories supporting the interaction with the 3D objects. In our test bed, two types of visualization devices have been considered: the main visualization terminal and a second screen based on a tablet pc.

The main visualization terminal is in charge of displaying the stereoscopic video stream and of playing the binaural sound in a synchronized manner. With this purpose, the MPEG-TS stream injected into the network by the receiver can be decoded directly and visualized by means of any video player compatible with the AVC video format. However, merging 3D objects in the video frame introduces an intermediate step, which is required to extract the depth information from the stereo image, compute the correct perspective (and occlusions) corresponding to the position of the 3D object within the scene, and generate the new stereo pair to be displayed with the 3D object properly embedded within it. The work by Galloso et al. [2012] describes the software algorithm implemented in our test bed for merging the interactive 3D objects into the video stream. The general idea behind this algorithm is to generate a displacement model of the video scene using depth maps, and to compute object occlusions attending to the Z-buffer information of both the scene and of the synthetic objects to be merged.

Fig. 6. Depth map results from stereoscopic images using the implemented CUDA SGBM algorithm with a min disparity of -12 pixels and 16 disparity levels. Results have been achieved using an Intel Core i7 - 3630QM processor and a NVIDIA GeForce GT 640M GPU (2 GB RAM dedicated). Video produced by and courtesy of Eumovil S.L.

Furthermore, the task of generating depth maps from stereoscopic video frames concerns another well-known problem in computer vision. To tackle this issue we rely on a semi-global block matching (SGBM) algorithm [Hirschmüller 2008] implemented on GPU using the Compute Unified Device Architecture (CUDA) tools provided by nVidia [Kowalczuk et al. 2012]. This method allows all the processing to be performed within the time imposed by the source video frame rate (25fps) thus preventing unwanted performance dropdowns. As shown in Fig. 6, the results obtained in our tesbed demonstrate that for sports event videos, the SGBM algorithm programmed in CUDA achieves an overall performance of 26.5 fps.

On the other hand, the secondary screen terminal (Fig. 7) has to access the web server content using the Internet communication channel. In this case, all the data are downloaded from the HTTP server on the producer side and are visualized using an HTML5 compatible web browser (e.g., Google

Chrome). To retain synchronism with the main screen, the downloaded web page includes a *javascript* socket script that connects to the local receiver terminal and periodically queries for the extracted PCR samples of the video.

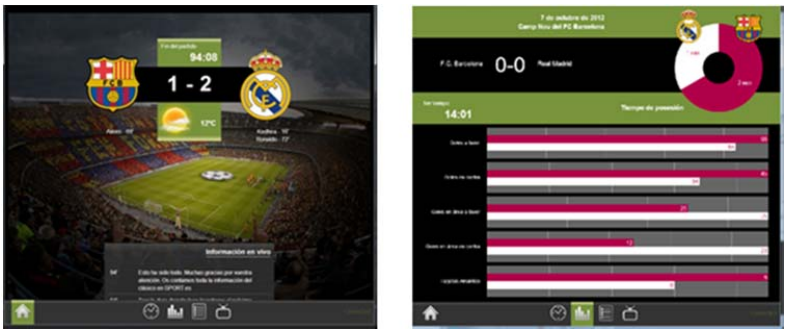

Fig. 7. Additional content displayed on the secondary screen using an HTML5 web browser

Finally, the set of devices used to manage the multi-sensorial environment are connected through a separate data bus according to the aforementioned KNX standard. The interconnection between the LAN and KNX networks is done through an IP/KNX router device that adapts the transmitted data to the physical layer requirements of each network. The task of programming and deploying this network in the room has been achieved using the ETS4 authoring tool, provided by the KNX association for the configuration of this specific technology. Although KNX has been the control protocol used in our test bed, it is important to state that the defined architecture could be adapted easily and scaled to integrate other actuators, regardless of their type, protocol technology, or the number of devices connected.

Fig. 8 shows the elements integrated in our test bed setup. The actuator devices shown in the right picture are programmed to scan the KNX bus to identify specific frames (identified by their physical address) encapsulating the action to be executed. In the architecture, these frames are generated by the scene control module, which is in charge of translating the user datagram protocol scene messages into valid KNX frames. The encoded list of devices/actions corresponding to each scene has to be made accessible from the scene control module according to the specifications listed in Appendix A.4.

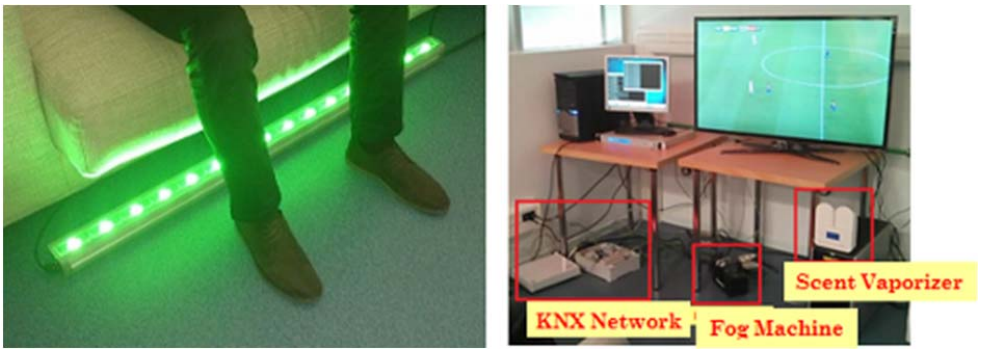

Fig. 8. Test bed setup: ambient light (left); KNX network, fog machine and scent vaporizer (right)

5. THE USER EXPERIENCE

An experimental study was conducted to study the perceived quality of the multi-sensorial stimuli and multi-modal interaction components introduced and their influence on the QoE. The experiment also

explored possible interactions between these elements at a global level and across the immersion and interaction dimensions.

5.1 Test environment and equipment

The test bed described in section 4 was deployed in a laboratory space recreating a standard living room. The following hardware and software components were used (number of units in brackets):

- Control station: (1) Intel Xeon W3565 3.2 GHz, 6 GB RAM, Graphic card: NVIDIA GeForce GTX 580 1536 MB GDDR5, Windows 7 (64 bits) + 3DVIA Virtools 5.0
- Transmitter: (1)PROMAX MO-170 DVB-T; (1)DTU-225 FantASI USB-2 ASI/SDI Input Adapter (DekTec); (1)Intel Xeon 3.2GHz, 2GB RAM, NVIDIA Quadro FX3000 256MB, Windows7 32 bits
- Receiver: (1) Intel Core2 Duo CPU E7300 2.66 GHz 2 GB RAM, ASUS My Cinema DVB-T Card, Debian GNU/Linux 6.0. Kernel release: 2.6.32
- Main screen: (1) Samsung SmartTV 46" UE46F5700AW; (2) active stereoscopic glasses
- Second screen: (1) iPad 2 GSM (A1396) 16GB
- Accessories: (1) Joystick Logitech RumblePad 2; (1) Headsets Sennheiser HD 545
- Sensory devices: (1) Olfactive Spirit Pro perfume diffuser (fragrance cartridge: CUT GRASS); (1) PHILIPS LEDline^2 BCS722 48xLED-LP/WW 6;
- KNX devices: (1) KNX/IP Router REG-K; (1) KNX/DALI Gateway REG-K/1/16(64)/64; (1) KNX Blind Actuator REG-K/4x/10

The LEDline unit includes 16 triads of non-dimmable low power RGB LED units resulting in up to 8 additive RGB colors (color depth = 3). It was placed in the front side of the sofa, attached all along the rack band, just behind the observer's feet (see Fig. 8 left). The scent vaporizer was placed near the bottom-right corner of the TV screen at a distance of 3.2m from the observer in the horizontal plane and at a height of 0.38m from the floor (see Fig. 8 right). Once the scent is emitted, it naturally fades until the end of the session when the blinds are automatically opened and the room is ventilated during 15 minutes (one session corresponds to one experimental condition (EC) as defined in Section 5.4). Although the fog machine was initially part of our test bed, we decided to exclude it from the QoE experiments after some preliminary tests where we found that the cut grass scent was heavily distorted by the smell of the smoke solution. The viewing distance and angle were equal to 3.5m and under 30º respectively as recommended in [ITU-R 2012]. All the experiments were conducted under the same ambient conditions and avoiding any interruption.

5.2 Content

Using the content described in Sections 4.1 and 4.2, we produced two video sequences (durations of 11'30" and 6'10", respectively) showing relevant scenes from both the first and second halves of the football match. In both sequences, the sensory effects and interaction components introduced (see Section 5.4 for a detailed description of those included under each EC) were designed carefully to complement the relevant events occurring during the game (see Table 1 and Appendices A.1 and A.4); thus, their frequencies and durations were in synchronization with those of these events. To avoid possible variations produced by the influence of different genres in the quality indicators under study, we used these two test sequences as unique audiovisual content and presented them in the same order to all the participants to ensure an experience as natural as possible.

5.3 Inclusion criteria

All the participants were non-experts and none was involved directly in the work under evaluation. Prior to the session, they were checked for having (corrected-to) normal visual acuity and stereopsis using a Snellen chart and self-developed materials based on Annex 1 of the ITU-R Recommendation BT.1438 [ITU-R 2000], respectively. Concerning their emotional and physical condition, all individuals were required to have a history free of neurological disease, head injury, learning

disability, and mental or psychological disorder, not to be using any medication for psychological or emotional problems, and not to be using any drug that could affect cognitive capabilities. The fulfillment of these requirements was checked using a self-developed screening questionnaire. The participants were also required to score lower than 18 on the Beck Depression Inventory [Beck 1961].

5.4 Procedure

We used a self-developed procedure inspired by the ITU-T Recommendation P.911 [ITU-T 1998], ITU-R Recommendations BT.1438 [ITU-R 2000], BT.500 [ITU-R 2012], and by previous research on QoE (e.g., see Jumisko-Pyykkö [2011]). To analyze the impact of each stimuli/component in an independent and combined way across the two dimensions considered, we manipulated the following variables:

Dimension 1 - Immersion; three immersive conditions (IC) are considered:
- BA. The binaural audio headsets are used to deliver high-quality spatial audio effects. No sensory effects are produced (actuator devices deactivated);
- SE. The actuator devices are activated and sensory effects are rendered in synchronization with the main audiovisual content. Standard stereo audio is delivered using the integrated loudspeakers of the TV equipment;
- BA+SE. Both the binaural audio headsets are used to deliver high-quality spatial audio effects, and actuator devices are activated and sensory effects rendered in synchronization with the main audiovisual content.

Dimension 2 – Interaction; three levels of interaction (Int) are defined:
- I3D. The interaction with 3D objects inserted into the main video scene is enabled. No tablet device is provided;
- ITab. Additional content can be accessed on demand through a tablet device. 3D objects are disabled;
- I3D+ITab. Both the interaction with 3D objects inserted into the main video scene and access on demand to additional content through a tablet device are enabled.

The resulting map of experimental conditions (ECs) is shown in Table 2.

Table 2 Map of experimental conditions

IC / Int	**I3D**	**ITab**	**I3D+ITab**
BA	EC1	EC2	EC3
SE	EC4	EC5	EC6
BA+SE	EC7	EC8	EC9

Two consecutive ECs were assigned randomly to each participant for the first and second test sequence, respectively (that is, we make a random assignment within-subjects and between-subjects of the immersion and interaction choices). During the session, the participant was invited to sit on the sofa and the following steps were implemented:

1. We explained in a quick and friendly manner the general goals of the experiment, procedure to be followed, and functionalities and principles of use of the interactive components presented. The explanation was adjusted to the specific ECs assigned to the participant (e.g., we only demonstrated how additional content could be accessed using the tablet if this component was presented in one of the ECs assigned to that participant). The participant was invited to manipulate the interaction components at will and to ask any question they had, and any issues raised were clarified;
2. Before launching the first test sequence, the user was invited to enjoy the experience freely, interacting (or not) at will with those elements presented in the assigned EC;

3. Immediately following the end of the test sequence, the participant was asked to rate the quality of each element presented: audio (QoA), sensory effects (QoSE), interaction with 3D objects (QoI3D), and interaction with tablet (QoITab); and the overall QoE, using a numerical scale from 0 (extremely bad) to 10 (excellent).
4. Finally, the second sequence was launched and step 3 repeated on its conclusion.

5.5 Results

Forty-eight individuals were recruited for the study: 35 males and 13 females. The participants were aged between 18 and 54 years (X = 29.6, SD = 8.59) and the sample comprised undergraduate students (37%), postgraduate researchers (50%), and post-doc researchers and university professors (13%). Thus, we achieved a total of 96 samples for all the ECs. Two outliers were discarded as recommended in ITU-R BT.500 [ITU-R 2002].

	MOS	SD	N
EC1	8.00	1.10	7
EC2	8.10	0.99	10
EC3	7.83	1.03	12
EC4	7.75	1.06	13
EC5	7.64	1.50	11
EC6	7.70	1.25	10
EC7	7.57	1.27	7
EC8	7.77	1.24	13
EC9	8.00	1.00	11

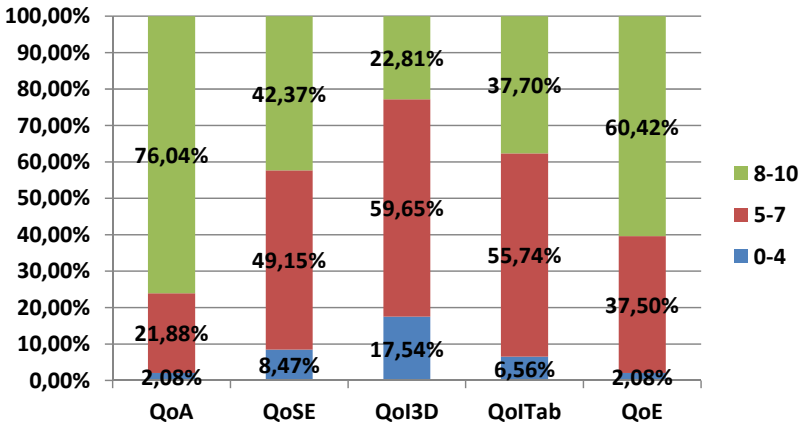

Fig. 9. (Left) MOS and SD for the QoE and number of individuals assigned to each EC; (Right) Percentage of quality assessment rates for each independent component and for the overall QoE

Fig. 9 (Left) shows the mean opinion score (MOS) and standard deviation (SD) for the overall QoE scores as well as the number of individuals (N) assigned to each EC (X = 10.44, SD = 2.24). As shown, the overall QoE was regarded as good in all of the cases. The higher values of MOS resulted for EC2 (BA, ITab), EC9 (BA+SE, I3D+ITab), and EC1 (BA, I3D), respectively. All these ECs present BA as a common immersive element, and the first two include ITab.

Fig. 9 (Right) shows the resulting quality assessments for all the components presented as a percentage from extremely to slightly bad (0–4), neutral to fair (5–7), and good to excellent (8–10). Audio was the component that received the best evaluation with 76.04% of assessments classed as good to excellent. This was followed by sensory effects, which were rated as good to excellent in 42.37% and as neutral to fair in 49.15% of the samples, respectively. Globally, the component that received the worst evaluation was the interaction with 3D objects; a significant 17.54% of assessments rated it as extremely to slightly bad and only 22.18% as good to excellent. This low scoring could be partially caused by the manner in which we approached this interaction component in the experiments, i.e., with the lack of concrete goals and/or rules/guidelines for the interaction with 3D objects. Finally, the overall QoE was perceived as good to excellent in 60.42% and as neutral to fair in 37.5% of the samples. The subjective assessment for the overall QoE was generally higher than that corresponding to each individual factor within the same sample, except for the audio quality, which was rated equally or higher than the QoE in 77% of the samples. The distribution of QoE and QoSE and of QoA, QoI3D and QoITab scores is normal and approximately normal, respectively.

We adjusted a *generalized linear model* to relate the perceived quality of each component and of the overall QoE with the four elements manipulated in the experiment (BA, SE, I3D and ITab), both across each dimension and for the overall sample. The numerical results of these analyses are provided in Appendix B.2 and summarized as follows.

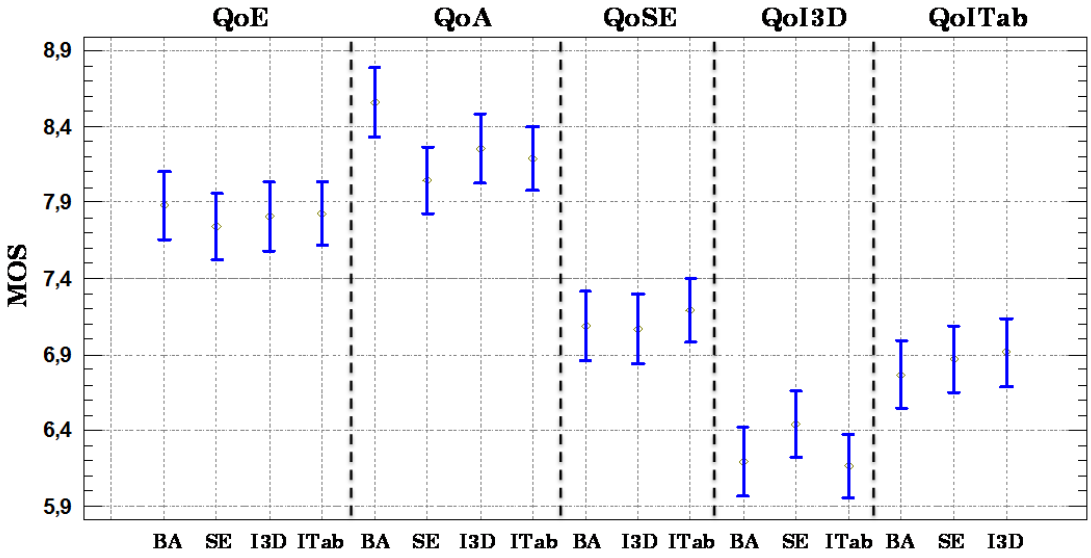

Fig. 10. MOS and 95% Fisher's LSD confidence interval for the subjective quality assessments across the two dimensions considered: immersion (BA and SE) and interaction (I3D and ITab)

Fig. 10 depicts the MOS and 95% Fisher's least significant difference (LSD) confidence interval of the quality assessments for each individual component and for the QoE across the two dimensions considered (immersion and interaction). As it shows, the MOS for QoA is on average higher than the values computed for the other components (QoSE, QoI3D and QoITab). In coherence with previous studies focused on this component, we found a very strong influence of BA in the perceived QoA, both across dimension 1 ($p=0.0098$) and for all dimensions ($p=0.0083$). Fig. 10 also suggests a slight influence of ITab on the QoSE but not enough statistical significance was achieved by the component ($p=0.2974$). Concerning interaction, we found that SE influence the QoI3D strongly across dimension 1 ($p=0.0464$) and moderately at a global level ($p=0.0531$). The relatively low p-value computed for ITab across dimension 2 ($p=0.2502$) also suggests a possible impact of the component on the QoI3D. No significant effects on the QoE were found. Globally, these initial models explain the variability of the QoA assessments in an 11.09% and of the QoI3D assessments in a 5.76%, which are mainly given by the influence of BA and SE factors, respectively.

Finally, we adjusted a *generalized linear model* to relate the overall QoE with the perceived quality of each component (see Appendix B.3). We found a very strong impact of the perceived QoSE ($p=0.0002$) and QoA (0.0004) on the QoE assessments, and a weak influence -potentially not significant- of the perceived QoI3D ($p=0.1553$). The simplified model, computed as a function of the two significant indicators, explains a 36.66% of the variability of the overall QoE assessments.

6. CONCLUSIONS AND FUTURE WORK

We have implemented an end-to-end solution integrating sensory effects and interactive elements into a hybrid (internet-broadcast) television system. Our system is fully compatible with current standards for transmission (DVB-T), video coding (AVC), data multiplexing (MPEG-2), additional signaling (DVB), and actuator devices (KNX). Furthermore, it can be easily adapted to support MPEG-V by

defining suitable *Sensory Effect Metadata* according to the *Sensory Effect Description Language* (SEDL) and *Sensory Effect Vocabulary* (SEV) specified by the standard, and by performing configuration adjustments in the receiver gateway to generate the KNX commands accordingly. Even though our solution has been tested with a sports multimedia asset, its implementation is completely transparent to the genre or format of the main audiovisual content.

However, even when our end-to-end fully functional test bed succeeded in demonstrating the technical feasibility of this approach, further important research and implementation challenges remain. For the content creation and adaptation phases, further efforts are required to enhance state-of-the-art effect authoring tools (see for example the work by Kim [2013] and references therein) with automatic annotation solutions. A possible approach could be applying semantic video annotation techniques to identify in an efficient and effective way relevant events that should trigger the sensory effects and/or interactive actions. Computer vision algorithms focusing on the recognition of specific scene features, objects or elements, as well as on human identification, could be also used to boost the visualization of additional content associated with the recognized element/character. Similarly, as pointed out by Timmerer et al. [2012], the (semi-) automatic generation of sensory effects presents several challenges that so far, have been only partially addressed (see, for example, the work in Waltl et al. [2009] with regard to the automatic extraction of color information from the video frame and the use of such information to control ambient lighting). In this sense, the automatic identification of elements in the video scene could also be used for the release of scents stimulating the olfactory sense in coherence with the presented content. A common implementation issue of these techniques is the necessity of decreasing the cost-intensity of the algorithms to enhance their deployment feasibility.

From a user perspective, audio was the element that received the best evaluation, followed at a considerable distance by sensory effects, on-demand interaction with additional content using the tablet device, and interaction with 3D objects as the least valuated. Surprisingly, during the experiments the participants generally showed themselves as enthusiastic about the interaction with 3D objects. Thus, we suspect that the low scoring of this last element could be partially due to the way in which we approached this interaction component in the experiments, i.e., with the lack of concrete goals and/or rules/guidelines. New improved experiments are deemed necessary to understand better the potential benefits and drawbacks of this interaction component.

We adjusted a *generalized linear model* to relate the perceived quality of each component and of the overall QoE with the four components introduced, both across each dimension and for the overall sample. The variability of QoA and QoI3D assessments was mainly given by the BA and SE components, respectively. A weak –potentially not significant- impact of the ITab component on the perceived QoSE and QoI3D was also observed. However, further studies are required to determine the nature and magnitude of such influence (if any), and how it could be bounded by other factors. In contrast to previous studies, no significant direct relationships on the QoE were found. To gain insight into this result, we analyzed the impact of the perceived quality of each component on the overall QoE. A strong influence of the QoSE and QoA was found. The simplified model, computed as a function of these two indicators explains the variability of the overall QoE assessments in a 36.66%. From here, further efforts would be required to increase the explicative power of our models, both enhancing their granularity (e.g. by decomposing the impact of each individual sensory effect), and considering the influence of other factors of technical, contextual and/or human nature (e.g. preferences and skills in relation to the content and components presented).

From a market deployment perspective, our work displays the potential of both immersion and interaction in video services and explores new services and content categories. 3D binaural audio and sensory effects show great potential for creating added-value services enabling immersion beyond stereoscopic video, in particular, for specific genres as documentaries, action movies, and sports. Interaction with 3D objects opens up the possibility to support interactive advertising or gaming applications, while on-demand interaction with additional content through a second screen seems to provide a suitable framework for personalized services and content.

ACKNOWLEDGEMENTS

This work has been developed partially within the framework of the ImmersiveTV project. The authors would like to thank all the participating entities and the Spanish Ministry of Industry Tourism and Commerce for its support through funded project TSI-020302-2010-61.

REFERENCES

Beck, A.T., 1961. An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6), p.561.

Bracken, C., Pettey, G. and Wu, M., 2011. Telepresence and Attention: Secondary Task Reaction Time and Media Form P. Turner, ed. *Proceedings of the International Society for Presence Research Annual Conference*.

Le Callet, P., Möller, S. and Perkis, A. EDS., 2012. Qualinet White Paper on Definitions of Quality of Experience.

Cha, J., Eid, M. and Saddik, A. El, 2009. Touchable 3D video system. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 5(4), pp.1–25.

Cobos, M., Lopez, J.J., Navarro, J.M. and Ramos, G., 2013. Subjective quality assessment of multichannel audio accompanied with video in representative broadcasting genres. *Multimedia Systems*.

ETSI, 2006. Digital Video Broadcasting (DVB); Multimedia Home Platform (MHP). European Telecomunications Standards Institute (ETSI) Technical Specification TS 101 812 v1.0.2.

ETSI, 2010a. Digital Video Broadcasting (DVB); DVB-HN (Home Network) Reference Model Phase 1. European Telecomunications Standards Institute (ETSI) Technical Specification TS 102 905 Ver. 1.1.1.

ETSI, 2010b. Digital Video Broadcasting (DVB); Signalling and carriage of interactive applications and services in Hybrid broadcast/broadband environments. European Telecomunications Standards Institute (ETSI) Technical Specification TS 102 809 v1.1.1.

ETSI, 2012a. Digital Video Broadcasting (DVB); Plano-stereoscopic 3DTV; Part 2: Frame Compatible Plano-stereoscopic 3DTV . European Telecomunications Standards Institute (ETSI) Technical Specification TS 101 547-2 V1.2.1.

ETSI, 2012b. HbbTV (Hybrid Broadcast Broadband TV) Technical Specification. European Telecomunications Standards Institute (ETSI) Technical Specification TS 102 796 v1.2.1.

Galloso, I., Luque, F.P., Piovano, L., Garrido, D., Sánchez, E. and Feijoo, C.A., 2012. Foundations of a New Interaction Paradigm for Immersive 3D Multimedia. In *Proceedings of 2012 NEM Summit: Implementing Future Media Internet towards New Horizons*. Istanbul, Turkey.

Ghinea, G. and Ademoye, O., 2012. The sweet smell of success: Enhancing multimedia applications with olfaction. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 8(1), pp.1–17.

Goldstein, E.B., 2010. *Sensation and Perception* C. Learning, ed.,

Heightman, D.W., 1975. Digital Video Broadcasting (DVB); DVB-HN (Home Network) Reference Model Phase 1. European Telecomunications Standards Institute (ETSI) Technical Specification TS 102 905 Ver. 1.1.1. *The Radio and Electronic Engineer*, 45(10), pp.559–569.

Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information. *IEEE transactions on pattern analysis and machine intelligence*, 30(2), pp.328–41.

INSPIRA, 2006. Informe sobre el uso de la televisión digital interactiva por parte del público español. Coordinated by G@TV-UPM. Sponsored by Spanish Ministry of Industry, Tourism and Trade.

ISO/IEC, 1998. Information technology — Coding of multimedia and hypermedia information — Part 6: Support for enhanced interactive applications. Organization for Standardization. International standard (ISO/IEC 13522-6).

ISO/IEC, 2000. Generic coding of moving pictures and associated audio information: Systems. International Organization for Standardization. International Standard (ISO/IEC 13818-1).

ISO/IEC, 2007. Home electronic system (HES) architecture Part 3-5. International Organization for Standardization. International standard (ISO/IEC 14543-3).

ISO/IEC, 2013. Information technology — Media context and control — Part 3: Sensory Information. Organization for Standardization. International standard (ISO/IEC 23005-3).

ITU-R, 2000. Subjective Assessment of Stereoscopic Television Pictures. *International Telecommunication Union*.

ITU-R, 2002. 500-11. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*.

ITU-R, 2012. Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service. *International Telecommunication Union*.

ITU-T, 1998. P.911 : Subjective audiovisual quality assessment methods for multimedia applications. *International Telecommunication Union*.

Jones, B.R., Benko, H., Ofek, E. and Wilson, A.D., 2013. IllumiRoom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. New York, New York, USA: ACM Press, p. 869.

Jumisko-Pyykkö, S., 2011. *User-Centered Quality of Experience and Its Evaluation Methods for Mobile Television*. Tampere University of Technology.

Kowalczuk, J., Psota, E. and Perez, L., 2012. Real-time Stereo Matching on CUDA using an Iterative Refinement Method for Adaptive Support-Weight Correspondences. , 23(1), pp.94–104.

Mills, P., Sheikh, A., Thomas, G. and Debenham, P., 2011. BBC R&D White Paper WHP208. , (December).

Murray, N., Qiao, Y., Lee, B., Karunakar, a. K. and Muntean, G.-M., 2013. Subjective evaluation of olfactory and visual media synchronization. *Proceedings of the 4th ACM Multimedia Systems Conference on - MMSys '13*, pp.162–171.

Nabi, R.L. and Krcmar, M., 2004. Conceptualizing Media Enjoyment as Attitude: Implications for Mass Media Effects Research. *Communication Theory*, 14(4), pp.288–310.

Olsen, D.R., Bunn, D., Boulter, T. and Walz, R., 2012. Interactive television news. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 8(2), pp.1–20.

Ramos, G. and Cobos, M., 2013. Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications. *The Journal of the Acoustical Society of America*, 134(3), pp.1735–8.

Raney, A.A. and Bryant, J., 2002. Moral Judgment and Crime Drama: An Integrated Theory of Enjoyment. *Journal of Communication*, 52(2), pp.402–415.

De Ruyter, B. and Aarts, E., 2004. Ambient intelligence. In *Proceedings of the working conference on Advanced visual interfaces - AVI '04*. New York, New York, USA: ACM Press, p. 203.

Schuemie, M.J., van der Straaten, P., Krijn, M. and van der Mast, C. a, 2001. Research on presence in virtual reality: a survey. *Cyberpsychology & behavior: the impact of the Internet, multimedia and virtual reality on behavior and society*, 4(2), pp.183–201.

Schulzrinne, H., Casner, S., Frederick, R. and Jacobson, V., 2003. RTP: A Transport Protocol for Real-Time Applications. *Internet Engineering Task Force, RFC 3550*.

Slater, M. and Wilbur, S., 1997. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence-Teleoperators and Virtual Environments*, 6(6), pp.603–616.

Slater, M.D., 2003. Alienation, Aggression, and Sensation Seeking as Predictors of Adolescent Use of Violent Film, Computer, and Website Content. *Journal of Communication*, 53(1), pp.105–121.

Timmerer, C., Waltl, M., Rainer, B. and Hellwagner, H., 2012. Assessing the quality of sensory experience for multimedia presentations. *Signal Processing: Image Communication*, 27(8), pp.909–916.

Waltl, M., Rainer, B., Timmerer, C. and Hellwagner, H., 2013. An end-to-end tool chain for Sensory Experience based on MPEG-V. *Signal Processing: Image Communication*, 28(2), pp.136–150.

Waltl, M., Timmerer, C. and Hellwagner, H., 2009. A test-bed for quality of multimedia experience evaluation of Sensory Effects. In *2009 International Workshop on Quality of Multimedia Experience*. IEEE, pp. 145–150.

Waltl, M., Timmerer, C. and Hellwagner, H., 2010. Increasing the user experience of multimedia presentations with sensory effects. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*. Desenzano del Garda: IEEE, pp. 1–4.

Wechsung, I., Schulz, M., Engelbrecht, K.-P., Niemann, J. and Möller, S., 2011. All users are (not) equal - The influence of user characteristics on perceived quality, modality choice and performance. In R. L.-C. Delgado & T. Kobayashi, eds. *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*. New York, NY: Springer New York, pp. 175–186.

Weffers-Albu, A., de Waele, S., Hoogenstraaten, W. and Kwisthout, C., 2011. Immersive TV viewing with advanced Ambilight. *2011 IEEE International Conference on Consumer Electronics (ICCE)*, pp.753–754.

Winer, D., 1999. XML Remote Procedure Call (XML-RPC) specification.

Yang, Z., Wu, W., Nahrstedt, K., Kurillo, G. and Bajcsy, R., 2010. Enabling multi-party 3D tele-immersive environments with ViewCast. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 6(2), pp.1–30.

Yoon, K., 2013. End-to-end framework for 4-D broadcasting based on MPEG-V standard. *Signal Processing: Image Communication*, 28(2), pp.127–135.

Zillmann, D. and Vorderer, P., 2000. *Media Entertainment: The Psychology of Its Appeal*, Lawrence Erlbaum Associates Publishers.

Online Appendix to: Integration of multi-sensorial stimuli and multi-modal interaction in a hybrid 3DTV system

FRANCISCO PEDRO LUQUE, IRIS GALLOSO AND CLAUDIO FEIJOO, Center for Smart Environments and Energy Efficiency (CEDINT), Universidad Politécnica de Madrid
CARLOS ALBERTO MARTÍN AND GUILLERMO CISNEROS, Grupo de Aplicación de Telecomunicaciones Visuales (G@TV), Universidad Politécnica de Madrid

A. COMPLEMENTARY TABLES AND FIGURES

A.1. Features and requirements of each content category

The following table summarizes the type of content, synchronization requirements, and rendering/visualization terminal corresponding to each content category integrated in our test bed

Content category	**Type of content**	**Synchronization**	**Rendering/visualization terminal**
Main audiovisual content	*3D video and (standard) audio*	Reference	Main 3DTV screen (and integrated audio system)
Additional content - Immersion	*Binaural audio*	To be played in synchronization with the main 3D video content	Binaural headsets
	Sensory effects: control commands triggering sensory devices	To be rendered in synchronization with the main audiovisual content	Lighting and shutter controllers, electronic scent vaporizer
Additional content - Interaction	*3D computer generated objects:* control commands activating / deactivating a 3D object or triggering specific animations	Activation/deactivation in synchronization with the main audiovisual content Automatic animations in synchronization with the main audiovisual content Asynchronous response to user interaction	Main 3DTV screen
	Additional information: numerical data (eg. score board, playing time, statistics); text (eg. game events, comments), 2D audiovisual content (eg. replays)	The displayed data shall be updated regularly (e.g. playing time) or when triggered by an event (e.g. score board) The 2D video stream shall be made available in the shortest time possible after the triggering event and shall be visualized asynchronously and on demand	Second screen (i.e., tablet PC)

DOI:http://dx.doi.org/10.1145/0000000.0000000

A.2. Editing decision list for our test bed

Duration	Start time	End time	Description
00:00:13:13	00:00:00:00	00:00:13:13	Teams logo (FCB-RMA)
00:02:42:00	00:00:13:13	00:02:55:13	Field occupation with beauty camera (Fast Forward)
00:03:39:21	00:02:55:13	00:06:35:09	Teams lineup and Steady Camera
00:05:10:00	00:06:35:09	00:11:45:09	Match start
00:03:46:00	00:11:45:09	00:15:31:09	Goal (Khedira RM)
00:00:13:13	00:15:31:09	00:15:44:24	Teams logo (FCB-RMA)
00:02:20:00	00:15:44:24	00:18:04:24	Second part ending
00:02:10:13	00:18:04:24	00:20:15:12	Time-lapse during rest-time
00:03:22:00	00:20:15:12	00:23:37:12	Second Part Start
00:00:13:13	00:23:37:12	00:23:51:00	Teams logo (FCB-RMA)
00:02:32:09	00:23:51:00	00:26:23:09	Goal (Alexis-FCB)
00:02:00:00	00:26:23:09	00:28:23:09	Goal (Cristiano Ronaldo-RMA)
00:00:13:13	00:28:23:09	00:28:36:24	Teams logo (FCB-RMA)
00:01:29:10	00:28:36:24	00:30:06:09	Match ending
00:04:35:19	00:30:06:09	00:34:42:03	Field des-occupation with beauty camera
00:00:13:13	00:34:42:03	00:34:55:16	Teams logo (FCB-RMA)

A.3. Format conversion and delivery channels

The following table summarizes the steps for adapting the format of each type of additional content in the post-production phase and the delivery channel used in each case.

Type of additional content	Input content (Content creation)	Output (Postproduction)	Delivery channel
Interactive 3D Models	-Raw object geometry in a modeling format (e.g., 3DS) -Animations -Materials and textures	- NMO file (graphic engine format) with all the content packed in one file and the object correctly positioned and scaled.	Broadband - web repository through a URL
Commands to interact with the 3D models	-List of actions/events related to 3D objects (type, duration, etc.)	- Text files in XML format for each action/event. - Synchronization time references related to the video	Broadcast - integrated in transport stream signal
Commands to enable/disable the actuator devices	- Actuator devices frame messages (e.g., KNX frames) - Table of scenes (environmental setups) and related actions	- Scene configuration file with a correspondence to a set of KNX frames -Synchronization time references related to the video	Broadcast - integrated in transport stream signal

Second screen	Information on events, players, match stats, etc.	- HTML web page with embed additional content	Broadband - web server content
Replays from other camera angles for the second screen	720p DNxHD recording from different camera perspectives	- AVC video clips to be delivered to a second screen via Internet	Broadband - web repository

A.4. Combination of multi-sensorial stimuli associated to each sensorial scene

The following table summarizes the combination of multi-sensorial stimuli associated to each sensorial scene defined in our test bed.

Scene command / actuator device	**Room shutters**	**Room lights**	**Ambient lights**	**Scent vaporizer**	**Fog machine**
Initial	OPEN	ON	OFF	OFF	OFF
Start	CLOSE	OFF	ON	OFF	OFF
Grass	CLOSE	OFF	ON	ON	OFF
Gameplay	CLOSE	OFF	ON	OFF	OFF
Goal	CLOSE	OFF	ON	OFF	ON
End	OPEN	OFF	OFF	OFF	OFF

B. STATISTICAL ANALYSIS

B.1. Terms, acronyms and clarifications

DF Degrees of freedom

MSE Mean squared error

MAE Mean absolute error

MAPE Mean Absolute Percentage Error

ME Mean Error

MPE Mean Percentage Error

SS Sum of Squares

Other symbols:

*0.05<p<0.1: low presumption against neutral hypothesis
**0.01<p<0.05: strong presumption against neutral hypothesis
***p<0.001: very strong presumption against neutral hypothesis

B.2. Numerical results on the relationship between the subjective quality assessments and the components manipulated in the experiment (BA, SE, I3D and ITab)

The following Table presents the ANOVA and Type III SS results corresponding to the perceived quality of each component and to the QoE measurements. It shows the impact on these indicators of each of the four elements manipulated in the experiment (BA, SE, I3D and ITab), both across the two dimensions considered (immersion and interaction) and for the overall sample. The values of ANOVA are highlighted in black whereas the Type III SS results are presented just below (within the same cell), corresponding in order to the following components: BA, SE, I3D and ITab.

Quality measures	Dimension 1 (BA vs. SE)			Dimension 2 (I3D vs. ITab)			All dimensions		
	F	**p-value**	**R^2**	**F**	**p-value**	**R^2**	**F**	**p-value**	**R^2 / adj. R^2**
QoA	**5.15** 12.25 0.16 - -	**0.0076**** 0.0098 0.7653 - -	**10.27%**	**0.08** - - 0.03 0.17	**0.9188** - - 0.9048 0.7717	**0.19%**	**2.74** 7.28 0.12 0.00 0.63	**0.0334**** 0.0083*** 0.7284 0.9978 0.4300	**11.09%**
QoSE	**0.02** 0.02 - - -	**0.8892** 0.8892 - - -	**0.02%**	**0.57** - - 0.08 1.78	**0.5689** - - 0.8332 0.3101	**1.25%**	**0.4** 0.09 - 0.05 1.10	**0.7501** 0.7653 - 0.8258 0.2974	**1.35%**
QoI3D	**2.25** 0.18 4.08 - -	**0.1118*** 0.6685 0.0464** - -	**4.75%**	**1.34** - - - 1.34	**0.2502** - - - 0.2502	**1.45%**	**1.81** 0.27 3.84 - 0.95	**0.1506*** 0.6019 0.0531* - 0.3321	**5.76%**
QoITab	**0.4** 1.36 0.18 - -	**0.6720** 0.3787 0.7473 - -	**0.88%**	**0.31** - - 0.31 -	**0.5782** - - 0.5782 -	**0.34%**	**0.36** 0.75 0.10 0.27 -	**0.7854** 0.3904 0.7493 0.6020 -	**1.18%**
QoE	0.44 0.36 0.23 - -	0.6447 0.5991 0.6759 - -	**0.97%**	0.04 - - 0.04 0.11	0.9570 - - 0.8609 0.7699	**0.1%**	0.22 0.26 0.16 0.02 0.02	0.9253 0.6129 0.6872 0.8951 0.8841	**1%**

B.3. Numerical results on the relationship between the QoE and the perceived quality of each component (QoA, QoSE, QoI3D and QoITab)

Dependent variable: QoE
Number of categorical factors: 0
Number of quantitative factors: 4 (QoA, QoSE, QoI3D, QoITab)
R^2 = 38.0983%
R^2 (adjusted) = 35.3162%

ANOVA Table (QoE)

Source	**Sum of squares**	**DF**	**Mean squares**	**F ratio**	**F probability (p-value)**
Model	44,4048	4	11,1012	13,69	0,0000
Residual	72,1484	89	0,810656		
Total (Corr.)	116,553	93			

Type III sum of squares (QoE)

Source	Sum of squares	DF	Mean squares	F ratio	F probability (p-value)
QoI3D	1,66514	1	1,66514	2,05	0,1553
QoITab	0,0116435	1	0,0116435	0,01	0,9049
QoSE	12,3304	1	12,3304	15,21	0,0002
QoA	10,8696	1	10,8696	13,41	0,0004
Residual	72,1484	89	0,810656		
Total (Corr.)	116,553	93			

Analysis of residuals

	Estimated value
n	94
MSE	0,810656
MAE	0,707754
MAPE	9,60478
ME	6,33063E-16
MPE	-1,43029

The equation of the adjusted model is:

$$QoE = 2.84613 + 0.115343*QoI3D - 0.00863951*QoITab + 0.30206*QoSE + 0.263957*QoA$$

The simplified model, computed as a function of the two significant indicators QoSE and QoA is as follows:

Dependent variable: QoE
Number of categorical factors: 0
Number of quantitative factors: 2 (QoA, QoSE)
R^2 = 36.6619%
R^2 (adjusted) = 35.2699%

ANOVA Table (QoE)

Source	**Sum of squares**	**DF**	**Mean squares**	**F ratio**	**F probability (p-value)**
Model	42,7306	2	21,3653	26,34	0,0000
Residual	73,8225	91	0,811237		
Total (Corr.)	116,553	93			

Type III sum of squares

Source	Sum of squares	DF	Mean squares	F ratio	F probability (p-value)
QoSE	14,5964	1	14,5964	17,99	0,0001
QoA	13,6824	1	13,6824	16,87	0,0001
Residual	73,8225	91	0,811237		
Total (Corr.)	116,553	93			

Analysis of residuals

	Estimated value
n	94
MSE	0,811237
MAE	0,698253
MAPE	9,49699
ME	-8,22037E-16
MPE	-1,46143

The equation of the adjusted model is:

$$QoE = 3.17719 + 0.321203*QoSE + 0.287788*QoA$$